When combinations of humans and AI are useful: A systematic review and meta-analysis

Year: 2024 | Citations: 156 | Authors: Michelle Vaccaro, Abdullah Almaatouq, Thomas W. Malone

Abstract

Inspired by the increasing use of artificial intelligence (AI) to augment humans, researchers have studied human–AI systems involving different tasks, systems and populations. Despite such a large body of work, we lack a broad conceptual understanding of when combinations of humans and AI are better than either alone. Here we addressed this question by conducting a preregistered systematic review and meta-analysis of 106 experimental studies reporting 370 effect sizes. We searched an interdisciplinary set of databases (the Association for Computing Machinery Digital Library, the Web of Science and the Association for Information Systems eLibrary) for studies published between 1 January 2020 and 30 June 2023. Each study was required to include an original human-participants experiment that evaluated the performance of humans alone, AI alone and human–AI combinations. First, we found that, on average, human–AI combinations performed significantly worse than the best of humans or AI alone (Hedges' g = −0.23; 95% confidence interval, −0.39 to −0.07). Second, we found performance losses in tasks that involved making decisions and significantly greater gains in tasks that involved creating content. Finally, when humans outperformed AI alone, we found performance gains in the combination, but when AI outperformed humans alone, we found losses. Limitations of the evidence assessed here include possible publication bias and variations in the study designs analysed. Overall, these findings highlight the heterogeneity of the effects of human–AI collaboration and point to promising avenues for improving human–AI systems. Vaccaro et al. present a systematic review and meta-analysis of the performance of human–AI combinations, finding that on average, human–AI combinations performed significantly worse than the best of humans or AI alone. They also found performance losses in decision-making tasks and significantly greater gains in content creation tasks.